面向不平衡数据分类的 KFDA-Boosting 算法 *

王 来 [1]，樊重俊 [1†]，杨云鹏 [1]，袁光辉 [2a, 2b]

(1. 上海理工大学 管理学院, 上海 200093; 2. 上海财经大学 a. 信息管理与工程学院; b. 实验中心, 上海 200433)

摘 要： 数据分布的不平衡性和数据特征的非线性增加了分类的困难，特别是难以识别不平衡数据中的少数类，从而影响整体的分类效果。针对该问题，结合 KFDA（kernel fisher discriminant analysis）能有效提取样本非线性特征的特性和集成学习中 Boosting 算法的思想，提出了 KFDA-Boosting 算法。为了验证该算法对不平衡数据分类的有效性和优越性，以 G-mean 值、少数类的查准率与查全率作为分类效果的评价指标，选取了 UCI 中 10 个数据集测试 KFDA-Boosting 算法性能，并与支持向量机等六种分类算法进行对比实验。结果表明，对于不平衡数据分类，尤其是对不平衡度较大或呈非线性特征的数据，相比于其他分类算法，KFDA-Boosting 算法能有效地识别少数类，并且在整体上具有显著的分类效果和较好的稳定性。

关键词： 核费希尔判别分析；集成学习；不平衡数据；分类

中图分类号： TP301.6

KFDA-Boosting algorithm oriented to imbalanced data classification

Wang Lai[1], Fan Chongjun[1†], Yang Yunpeng[1], Yuan Guanghui[2a, 2b]

(1. *Business School*, *University of Shanghai for Science & Technology*, *Shanghai* 200093, *China*; 2. *a. School of Information Management & Engineering*, *b. Experimental Center*, *Shanghai University of Finance & Economics*, *Shanghai* 200433, *China*)

Abstract: The imbalance of data distribution and the nonlinearity of data characteristics increase the difficulty of classification, especially the recognition of the minority class samples in the imbalanced data, thus affecting the overall classification effect. For the above problem, an algorithm called KFDA-Boosting was proposed in this paper, which combined the characteristic of KFDA , namely Kernel Fisher Discriminant Analysis, effectively extracting the samples' nonlinear features and the idea of Boosting algorithm in the ensemble learning. In order to verify the effectiveness and superiority of the algorithm in the classification of imbalanced data, the paper used the G-mean value, the precision and recall of the minority class samples to evaluate the performance of classifier, and selected 10 datasets of UCI to test the KFDA-Boosting algorithm, which compared with other six algorithms, such as support vector machine. Compared with other algorithms, the results show that the algorithm can effectively identify the minority class, and has a significant effect on the classification of imbalanced data and better stability on the whole, especially for the data with larger unbalance degree or nonlinear characteristics.

Key Words: Kernel Fisher discriminant analysis; ensemble learning; imbalanced data; classify

0 引言

不平衡数据分类，在许多领域中有着重要的应用，如疾病诊断、文本识别、入侵检测等。所谓不平衡数据，即数据集中某一类或某些类样本数远多于其他类别。对于不平衡数据分类，人们更多地关注少数类样本，并且少数类的错分代价相对较大。同时，随着数据量的增加，数据间越来越呈现非线性的特征，甚至为强非线性，这也增加了识别少数类的困难。因此，有效利用样本特征，精准地识别少数类样本，从而改善整体的分类效果，对解决不平衡数据的分类问题具有重大的价值与意义。

近年来，针对不平衡数据分类问题的研究，研究人员主要是从数据层和算法层两个层面着手。

在数据层面上，主要是通过重采样实现各类样本数的平衡，其中包括欠采样和过采样。对于重采样方法的研究，主要围绕 Laurikkala 提出的邻域清除算法[1]和 Chawla 等人提出的 SMOTE 算法[2]展开。例如，郑文昌等人[3]提出了面向不平衡数

基金项目：国家自然科学基金资助项目(71303157)；上海市教育委员会科研创新重点基金项目(14ZZ131)；上海市一流学科资助基金项目(S1205YLXK)；上海市社科规划青年课题基金项目（2014EGL007）；沪江基金资助项目（D14008）

作者简介：王来（1992-），男，湖北黄冈人，硕士研究生，主要研究方向为优化算法、机器学习；樊重俊（1963-），男（通信作者），山西运城人，教授，博导，主要研究方向为信息系统工程（fan.chongjun@163.com）；杨云鹏（1991-），男，甘肃天水人，博士研究生，主要研究方向为智能优化算法、大数据挖掘、跨境电子商务；袁光辉（1988-），男，陕西西安人，博士研究生，主要研究方向为智能优化算法、投资组合风险.

据集的 SMOTE-SVM 交通事件检测算法，其中采用了 SMOTE 算法对事件数据进行过采样，以降低不平衡性。类似地，衣柏衡、杨毅等人[4, 5]先通过改进的 SMOTE 算法平衡各类别的样本数，再基于分类算法处理不平衡数据。但过采样可能会引入其他噪声，而欠采样可能丢失某些有用的重要信息。在算法层面上，主要包括代价敏感学习、集成学习等方法。代价敏感学习，为不同类型的错误分配不同的代价，以达到分类时产生的错误总代价最低的目标[6, 7]。例如，邹鹏等人[8]针对客户价值细分问题中的不平衡数据，设计了代价敏感决策树算法，以实现对客户价值的有效识别。师彦文等[9]针对不平衡数据集，提出了将代价敏感和随机森林相结合的分类算法。而集成学习，主要包括 Boosting 算法[10]和 Bagging 算法[11]。对于运用集成学习解决不平衡数据分类的研究，较多的是在 Freund 和 Schapire 两人提出的 Boosting 算法的基础上进行改进[12, 13, 14]。例如，应维云等人[15]将 LDA 加入 Boosting 算法中建立弱分类器，应用到客户流失预测中。虽然 LDA-Boosting 提高了分类效率，但对包含非线性特征的不平衡数据进行分类时，并不能达到理想的效果；李诒靖等人[16]以 KNN 作为弱分类器，利用 BPSO 对数据进行特征提取后采用 Adaboost-KNN 算法进行分类，但最优特征子集的选取容易陷入局部最优解，进而影响最终的分类效果。

以上两个层面，并未充分考虑到样本特征的有效利用，尤其是非线性特征。针对该问题，考虑到 Boosting 作为一种有效的分类学习方法，在处理那些难以学习的样本时会赋以更高的权重，使得分类器在下次训练中聚焦到那些样本上，从而能够在一定程度上提升对不平衡数据的分类效果。而核 Fisher 判别分析能够十分有效地对非线性特征进行提取，本文将这两种算法结合起来，提出了 KFDA-Boosting 算法。

KFDA-Boosting 算法利用核 Fisher 判别分析有效地提取非线性判别特征，并借助集成学习中 Boosting 算法的思想改善其分类性能。最后，对 UCI 中选取的 10 个数据集进行了仿真实验，以测试 KFDA-Boosting 算法对不平衡数据分类的可行性和有效性，并和其他六种分类算法的分类效果对比分析，期望体现该算法对少数类的有效识别，且整体的分类效果有一定的提升。

1　KFDA-Boosting 算法

1.1　Boosting 算法思想

本文主要以二分类问题为例，阐述 Boosting 算法思想[17, 18]如下：

给定弱分类器和训练集

$$S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$$

其中：$x_i(i=1,2,\cdots,m)$ 是一个 n 维列向量，y_i 表示第 i 个样本的标签，$y_i\in Y=\{+1,-1\}$。

首先，对训练集中各样本赋予一个初始权重。接着，在每轮迭代过程中，会产生一个弱假设 $h_t:X\to\{1,-1\}$，并相应地更新样本权重，即增大那些被错误分类的样本的权重，减小分类正确的样本的权重，使得弱分类器在下次迭代中集中到那些被错分的样本上。经过 T 轮迭代后，根据其分类精度得到的权重，将每轮迭代过程中产生的弱假设进行加权，从而得到最终的分类器。

1.2　改进的核 Fisher 判别分析

核 Fisher 判别分析的基本思想是将一输入空间 R 非线性映射到另一个特征空间 F，然后在特征空间中利用 Fisher 判别分析，以达到对输入空间进行分类的目的[19, 20]。

考虑到 Boosting 算法的特点，本文对原始的核 Fisher 判别分析加以改进，即对每个样本赋予相应的权重。具体思路与过程如下：

在特征空间 F 中，每个样本对应的权重为 $D_{t,i}$，其看做是第 i 个样本，在第 t 轮学习中的权重。在变换后的空间 F 中，各类样本的均值 $m_{t,k}^{\phi}$ 和样本类内离散度矩阵 $S_{t,k}^{\phi}$ 分别为

$$m_{t,k}^{\Phi}=\frac{\sum_{y_i=k}D_{t,i}\phi(x_i)}{\sum_{y_i=k}D_{t,i}},k=-1,1 \tag{1}$$

$$S_{t,k}^{\phi}=\frac{\sum_{y_i=k}D_{t,i}(\phi(x_i)-m_{t,k}^{\phi})(\phi(x_i)-m_{t,k}^{\phi})^T}{\sum_{y_i=k}D_{t,i}},k=\pm1 \tag{2}$$

由上述 $m_{t,k}^{\phi}$ 和 $S_{t,k}^{\phi}$，样本类间离散度矩阵 $S_{t,B}^{\phi}$ 和样本类内离散度矩阵 $S_{t,W}^{\phi}$ 可表示为

$$S_{t,B}^{\phi}=(m_{t,1}^{\phi}-m_{t,-1}^{\phi})(m_{t,1}^{\phi}-m_{t,-1}^{\phi})^T \tag{3}$$

$$S_{t,W}^{\phi}=\sum_{y_i=1}D_{t,i}S_{t,1}^{\phi}+\sum_{y_i=-1}D_{t,i}S_{t,-1}^{\phi} \tag{4}$$

即

$$S_{t,W}^{\phi}=\sum_{k=1,-1}\sum_{y_i=1}D_{t,i}\left(\phi(x)-m_{t,k}^{\phi}\right)\left(\phi(x)-m_{t,k}^{\phi}\right)^T \tag{5}$$

根据 Fisher 判别准则，此时 Fisher 准则函数的表达式为

$$J_F(w)=\frac{w_t^TS_{t,B}^{\phi}w_t}{w_t^TS_{t,W}^{\phi}w_t} \tag{6}$$

由此得到了特征空间中的 Fisher 判别函数，从而实现 Fisher 判别。但如果特征空间 F 维数非常高，甚至为无限维时，无法直接通过上式求解最佳判别矢量。针对该问题，引入核函数。本文采用 RBF 核函数（即 Gauss 径向基核函数）作为映射函数 ϕ，即

$$k(x,y)=\exp(-\frac{\|x-y\|^2}{\sigma^2}) \tag{7}$$

其中：x，y 为对应的样本值；σ 为常数，其决定非线性化的程度。

由再生核理论可知，高维空间中的任一解都可以被表示为该空间中训练样本线性组合的形式，即有

$$w_t=\sum_{i=1}^{m}\alpha_i\phi(x_i)=\alpha\phi(x) \tag{8}$$

其中：$\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_m)^T\in R^n$ 为各个元素 $\phi(x_i)$ 的线性系数。

根据 $m_{t,k}^{\phi}(k=-1,\ 1)$ 的定义和式(8)，将特征空间 F 中训练样

本的均值 $m_{t,k}^{\phi}$ 投影到 w_t 上，可得

$$w_t^T m_{t,k}^{\phi} = \frac{1}{\sum_{y_i=k} D_{t,i}} \sum_{j=1}^{m} \sum_{y_i=k} \alpha_j k(x_j, x_i^k) D_{t,i} = \alpha^T M_{t,k} \tag{9}$$

其中：

$$M_{t,k} = \frac{1}{\sum_{y_i=k} D_{t,i}} \left(\sum_{y_i=k} k(x_1,x_i), \sum_{y_i=k} k(x_2,x_i), \cdots, \sum_{y_i=k} k(x_m,x_i) \right) \tag{10}$$

则可将式(6)中右边分式的分子、分母分别转换为如下形式：

$$\begin{aligned} w_t^T S_{t,B}^{\phi} w_t &= w_t^T \left(m_{t,1}^{\phi} - m_{t,-1}^{\phi}\right)\left(m_{t,1}^{\phi} - m_{t,-1}^{\phi}\right)^T w_t \\ &= \alpha^T (M_{t,1} - M_{t,-1})(M_{t,1} - M_{t,-1})^T \alpha \\ &= \alpha^T M \alpha \end{aligned} \tag{11}$$

其中：

$$M = (M_{t,1} - M_{t,-1})(M_{t,1} - M_{t,-1})^T \tag{12}$$

$$\begin{aligned} w_t^T S_{t,W}^{\phi} w_t &= w_t^T \sum_{k=1,-1} \sum_{y_i=1} D_{t,i} \left(\phi(x) - m_{t,k}^{\phi}\right)\left(\phi(x) - m_{t,k}^{\phi}\right)^T w_t \\ &= \alpha_t^T H \alpha_t \end{aligned} \tag{13}$$

其中：

$$H = \sum_{k=1,-1} \sum_{y_i=k} D_{t,i} (k_{x_i} - M_{t,k})(k_{x_i} - M_{t,k})^T \tag{14}$$

$$k_{x_i} = \left(k(x_1,x_i), k(x_2,x_i), \cdots, k(x_m,x_i)\right)^T \tag{15}$$

联立式（6）（11）（13）可得，特征空间中的 Fisher 判别式转变为

$$J(\alpha) = \frac{\alpha^T M \alpha}{\alpha^T H \alpha} \tag{16}$$

根据广义的 Ralyeigh 熵的性质得

$$\alpha = H^{-1}(M_{t,1} - M_{t,-1}) \tag{17}$$

所以，特征空间 $\phi(x)$ 在 α 上的投影为

$$w_t \phi(x) = \sum_{j=1}^{m} \alpha_j k(x_j, x) \tag{18}$$

1.3 KFDA-Boosting 算法流程

将改进的核 Fisher 判别分析加入到 Boosting 算法框架之中，得到 KFDA-Boosting 算法流程如下。

对于训练集

$$S = \{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\},$$

其中：y_i 表示第 i 个样本的标签，$y_i \in Y = \{+1,-1\}$；弱学习算法为改进的 KFDA，迭代次数为 T，对于第 i 个样本在第 t 轮迭代时的分布为记为 $D_{t,i}$。

对于每轮迭代过程中的弱假设 $h_t : X \to \{1,-1\}$，其分类效果由错误率 ε_t 衡量：

$$\varepsilon_t = \sum D_{t,i} I[h_t(x_i) \neq y_i] = P_{D_{t,i}}(h_t(x_i) \neq y_i) \tag{19}$$

KFDA-Boosting 分类算法

输入：训练集 $S = \{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，其中 $x_i \in X$，$y_i \in Y = \{+1,-1\}$，迭代次数 T；

1 初始化样本权重：$D_{1,i} = \begin{cases} \frac{1}{2N_+}, \text{if } y_i = 1 \\ \frac{1}{2N_-}, \text{else} \end{cases}$，其中 N_+，N_- 分别为正类、负类样本的总数；

2 for t=1 to T

3 训练针对样本分布为 $D_{1,t}$ 的加权 KFDA

4 求解得到 $H, M_{t,k}, k = 1,-1$；

5 求解得到最优 α；

6 得到弱分类器 $h_t(x_i) = \begin{cases} +1, & \text{if } w_t\phi(x_i) > \theta_t \\ -1, & \text{else} \end{cases}$，阈值 θ_t 由各类样本均值在 α 上投影的加权平均值决定，权重分别为对应类的样本总数；

7 计算分类错误率 $\varepsilon_t = \sum_{k=1,-1} \sum_{y_i=k} D_t(i) I(h_t(x_i) \neq y_i)$

8 if $\varepsilon_t > 0.5$

9 continue

10 else if T=t-1

11 break

12 end if

13 令 $\alpha_t = \frac{1}{2} \log \frac{1-\varepsilon_t}{\varepsilon_t}$；

14 更新样本权重：for i=1 to m

15 $D_{t+1,i} = \frac{D_{t,i}}{Z_t} \exp(-\alpha_t (I(h_t(x_i) = y_i)))$，其中 Z_t 为归一化算子，使得 $\sum_i D_{t+1,i} = 1$；

16 end for i

17 end for t

输出：最终假设 $H(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$

2 分类评价指标

在不平衡数据中，少数类对应为正类，多数类对应为负类，表 1 给出了二分类问题的混淆矩阵。

表 1 二分类问题的混淆矩阵

	正类（预测）	负类（预测）
正类（实际）	TP	FN
负类（实际）	FP	TN

在传统的分类算法中，通常采用分类准确率作为分类性能评价指标，即

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

通常情况下，不平衡数据中的正类样本数所占比例小，因而 TP 不会太大，甚至出现为 0 的情况，而 TN 值很大，使得分类的最终准确率较大，但却因此忽略了分类器对正类的识别正确率。所以，准确率并不能真正意义上反映分类器的性能。

鉴于分类正确率存在以上缺陷，本文采用 G-mean 值作为整体的分类性能评价指标，即

$$G-mean=\sqrt{TPR\cdot TNR} \tag{21}$$

其中：$TPR=\frac{TP}{TP+FN}$，$TNR=\frac{TN}{TN+FP}$。

G-mean 值作为不平衡数据分类常用的评价指标，分别用 TPR、TNR 来衡量正类和负类的分类性能，其值越大表明分类效果越好。二者其中若有一个值的结果不佳，就会导致 G-mean 值不理想。

同时，为了进一步衡量对正类的分类效果，在此引入正类的查准率（precision）与查全率（recall），其定义分别如下：

$$precision=\frac{TP}{TP+FP} \tag{22}$$

$$recall=\frac{TP}{TP+FN} \tag{23}$$

3　算法实验与结果分析

3.1　数据来源

从 UCI 中选取了 10 个数据集作为测试数据。为了将所选的数据集看做二分类问题的研究对象，作以下规定：如果数据集为两类，则将其中数目较少的一类作为正类，如 Sonar、Ionosphere 数据集等；如果数据集为多类别，即其类别数大于 2，将其中的某一类作为正类，剩下的类统一合并当作负类。经过上述规定与处理后，按不平衡度（IL）大小升序排列，得到用于二分类的不平衡数据集情况，如表 2 所示。

3.2　数据预处理

为了防止属性值之间差距过大，而影响算法的迭代过程。因此，在进行算法实验之前，对原始数据进行归一化处理。

对于数据表中的任一特征属性，选取该特征属性数值中的最大取值，然后将所有样本的该属性值除以上述最大值得到各样本对应的归一化值。即计算公式如下所示：

$$x_{i,j}'=\frac{x_{i,j}}{\max(x_j)},i=1,\cdots,m;j=1,\cdots,N \tag{24}$$

其中：$\max(x_j)$ 表示样本第 j 个特征属性数值中的最大值。

表 2　实验数据集

数据集名	样本数	特征数	类别数	正类数/负类数	IL
Sonar	208	60	2	97/111	1.14
Ionosphere	351	34	2	126/225	1.79
Seeds	210	7	3	70/140	2.00
Wine	178	13	3	48/130	2.71
Ecoli	336	7	8	52/284	5.46
Fertility	100	9	2	12/88	7.33
Balance	625	4	3	49/576	11.76
Glass	214	9	7	13/201	15.46
Page-	5473	10	5	88/5385	61.19
Yeast	1484	8	10	20/1464	73.20

经过上述处理后，各样本的特征属性值转换为[0,1] 的数，同时这也消除了量纲的影响，且便于后续的迭代计算。

3.3　对比实验与结果分析

本文实验采用了五折交叉验证，通过随机地将数据集等分成五份，每次将其中的一份数据集将其中的作为测试集，另外四份则作为训练集。最后，将五次实验得到评价指标值（包括正确率、G-mean 值、少数类查准率与查全率）的平均值，即作为该算法测试的最终评价结果。其中，正确率虽然对于不平衡数据分类的效果评价上存在缺陷，但本文计算该值主要是用作对比说明。

为了验证 KFDA-Boosting 算法对不平衡数据分类的有效性，本文与其他六种算法进行对比实验，即决策树（DT）、支持向量机（SVM）、人工神经网络（ANN）、核 Fisher 判别分析（KFDA）、基于代价敏感的决策树（CS-DT）、结合过采样 SMOTE 算法的支持向量机（SMOTE-SVM）。其中 SVM、KFDA 与 KFDA-Boosting 使用的是同种核函数，即 RBF 核，且最大迭代次数设置为 200。另外，为了便于比较，其中的 SMOTE-SVM 和 CS-DT 参数设置方式分别同文献[3,8]。

通过实验得到以上七种算法的正确率、G-mean 值、正类查准率和查全率最终结果，其对比情况分别如表 3~6 所示。为了更加直观地比较分析各算法的分类效果，将表 3~6 中的测试结果绘制成对应的折线图，如图 1~4 所示。

表 3　各算法测试的正确率对比情况

编号	数据集名	DT	SVM	ANN	KFDA	CS-DT	SMOTE-SVM	KFDA-Boosting
1	Sonar	0.7200	0.8621	0.8196	0.8103	0.8516	**0.8856**	0.8276
2	Ionosphere	0.8429	0.8714	0.8429	0.9285	0.8757	**0.9247**	0.9428
3	Seeds	0.8535	0.9167	0.9167	0.9048	0.8975	**0.9375**	0.9286
4	Wine	0.9558	0.9470	0.9667	0.9485	0.9683	0.9653	**0.9874**
5	Ecoli	0.7264	0.9254	0.5475	0.6870	**0.9628**	0.9528	0.9491
6	Fertility	**1.0000**	0.8808	**1.0000**	0.8168	**1.0000**	0.9112	0.9625
7	Balance	0.5585	0.7546	0.8992	0.6960	0.8239	0.8967	**0.9040**
8	Glass	0.8926	0.9444	0.9302	0.7701	0.9153	**0.9513**	0.9247
9	Page-Blocks	0.9569	0.9508	0.9673	0.8537	**0.9571**	0.9281	0.9486
10	Yeast	0.9623	0.9822	**0.9865**	0.9833	0.9586	0.9845	0.9857

表 4　各算法测试的 G-mean 值对比情况

编号	数据集名	DT	SVM	ANN	KFDA	CS-DT	SMOTE-SVM	KFDA-Boosting
1	Sonar	0.5960	0.8610	0.6975	0.7542	0.8235	**0.8663**	0.8424
2	Ionosphere	0.8435	0.8919	0.7755	0.8512	0.8652	0.9111	**0.9271**
3	Seeds	0.8008	0.9106	0.9192	0.9250	0.8896	**0.9467**	0.9445
4	Wine	0.9252	0.9265	0.9542	0.9236	0.9578	0.9655	**0.9837**
5	Ecoli	0.7096	0.8706	0.6041	0.6996	**0.8875**	0.8766	0.8558
6	Fertility	**1.0000**	0.0000	**1.0000**	0.8885	**1.0000**	0.8589	0.9787
7	Balance	0.0000	0.4517	0.5638	0.7324	0.7853	0.8574	**0.8622**
8	Glass	0.8702	0.9014	0.9247	0.8732	0.8827	**0.9375**	0.9289
9	Page-Blocks	0.7806	0.8127	0.8806	0.8208	0.9264	0.9051	**0.9325**
10	Yeast	0.0984	0.5891	0.6228	0.8052	0.7153	0.7468	**0.9464**

表 5　各算法测试的正类查准率对比情况

编号	数据集名	DT	SVM	ANN	KFDA	CS-DT	SMOTE-SVM	KFDA-Boosting
1	Sonar	0.7188	0.8468	0.7941	0.8095	0.7425	**0.8451**	0.8293
2	Ionosphere	0.7586	0.7429	0.9412	0.8422	0.8530	0.8783	**0.9055**
3	Seeds	0.8162	0.8643	0.8467	0.7875	0.8817	**0.9034**	0.8889
4	Wine	0.9267	**1.0000**	0.9091	0.9546	0.9315	0.9785	0.9764
5	Ecoli	0.3043	0.7273	0.2051	0.2571	0.8914	0.9645	0.8876
6	Fertility	**1.0000**		**1.0000**	0.4805	**1.0000**	0.9216	0.9560
7	Balance	0.0000	0.0903	0.4570	0.1628	0.7563	0.8212	**0.8547**
8	Glass	0.5443	0.4536	0.6521	0.2404	0.8145	**0.8776**	0.8643
9	Page-Blocks	0.5455	0.5455	0.7619	0.7182	0.8457	0.8125	**0.8768**
10	Yeast	0.0265	0.5869	0.6733	0.6576	0.7542	0.8234	**0.8538**

表 6　各算法测试的正类查全率对比情况

编号	数据集名	DT	SVM	ANN	KFDA	CS-DT	SMOTE-SVM	KFDA-Boosting
1	Sonar	0.6216	0.8789	0.7297	0.9120	0.8134	**0.9268**	0.9189
2	Ionosphere	0.8462	0.9546	0.6154	**0.9873**	0.9673	0.9531	0.9857
3	Seeds	0.7857	0.8929	0.9286	**1.0000**	0.8913	0.9876	**1.0000**
4	Wine	**1.0000**	0.9169	**1.0000**	**1.0000**	**1.0000**	0.9543	**1.0000**
5	Ecoli	0.7236	0.8208	0.8000	0.9122	**0.9351**	0.9218	0.9265
6	Fertility	**1.0000**	0.0000	**1.0000**	**1.0000**	**1.0000**	0.8765	**1.0000**
7	Balance	0.0000	0.2592	0.3386	0.5778	0.7868	0.8427	**0.8905**
8	Glass	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
9	Page-Blocks	0.6777	0.6667	0.8889	0.7533	0.8562	**0.9414**	0.9353
10	Yeast	0.0575	0.5586	0.4563	0.7342	0.7896	0.8011	**0.9279**

由图 1 和 2 可以看出，与 DT、SVM、ANN 及两种改进的分类算法（CS-DT、SMOTE-SVM）相比，在表 3 中的五个数据集测试结果上，本文 KFDA-Boosting 算法的 G-mean 值最大，而在其他数据集上的 G-mean 值与对应的最大值相差并不大，这表明本文算法整体分类效果良好。当数据集的不平衡度逐渐增大时，与传统的 DT 和 SVM 算法相比，CS-DT 和 SMOTE-SVM 在个数据集上测试的 G-mean 值均有不同程度的增大，整体分类效果得到较好地改善。而在不平衡度增大一定程度时，本文算法仍具有很大的 G-mean 值，且相对优于 CS-DT 和 SMOTE-SVM 两种改进算法的对应值，如测试集 Page-Blocks、Yeast。与此同时，DT、SVM、ANN 三种算法测试的正确率虽均达到了 90%以上，但其对应的 G-mean 值却较小。

在图 3 与 4 中进一步可以看出，随着不平衡度逐渐增大，KFDA-Boosting 在对应数据集上测试的正类查准率与查全率两

项指标的均值都很大，而其他算法的对应值相对较小，该情况在呈非线性特征的不平衡数据测试上表现得尤为明显，例如数据集 Yeast，这表明本文算法能有效利用样本的非线性特征，且对少数类样本具有很强的识别能力。从侧面也证实了，分类正确率作为不平衡数据分类的评价指标有时并不能有效地衡量分类器的分类效果。

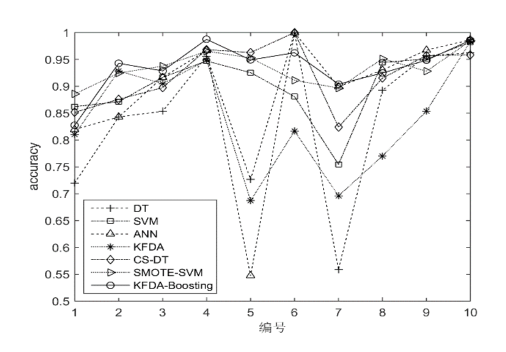


图 1　各算法测试的正确率对比

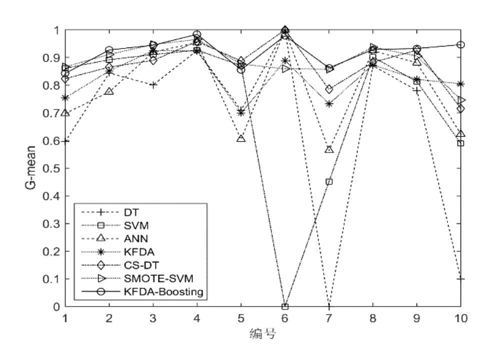


图 2　各算法测试的 G-mean 值对比图

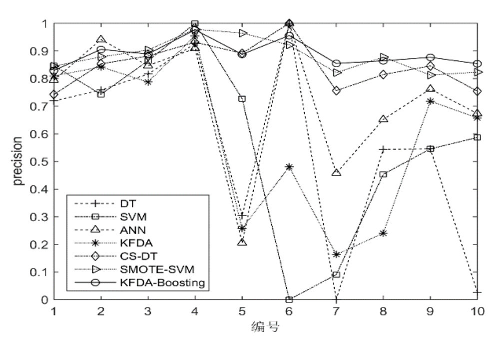


图 3　各算法测试的正类查准率对比

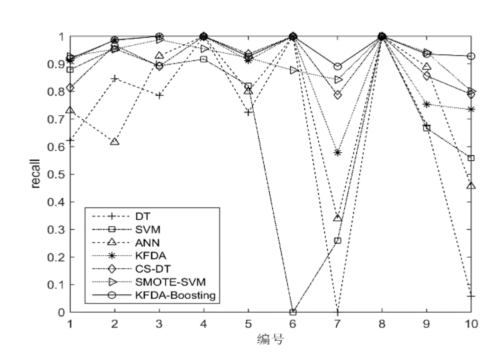


图 4　各算法测试的正类查全率对比图

同时还可以看出，对于同一数据集，KFDA-Boosting 算法的 G-mean 值及正类查准率大于 KFDA 对应的值。除数据集 Ionosphere 外，本文算法的正类查全率大于 KFDA 对应的指标值或与对应最优值相当，这表明本文提出的算法，与单独采用 KFDA 相比，分类效果有很大地提升，尤其是针对正类样本的识别。

此外，进一步可以算出各种分类算法测试的 G-mean 值方差分别为 0.1173、0.0889、0.0263、0.0063、0.0069、0.0039、0.0025，由此可说明，针对不同的数据集，本文提出的 KFDA-Boosting 算法与其他分类算法相比，整体分类具有较好的稳定性。

4　结束语

对于不平衡数据分类中，数据特征越来越呈现非线性，加大了识别少数类的困难，本文提出了一种基于改进的核 Fisher 判别分析与 Boosting 算法的分类方法，即 KFDA-Boosting 算法。该算法能有效利用样本特征，尤其是非线性特征，以实现对原始数据的非线性判别，保证了样本的最佳可分离性。最后，本文算法通过对 UCI 中的 10 个数据集的测试实验表明，对于不平衡度较大或呈非线性特征的数据，本文算法分类的效果显著，与 DT、SVM、ANN、KFDA、CS-DT、SMOTE-SVM 相比，KFDA-Boosting 算法能有效地识别少数类，表现出良好的整体分类效果，并具有较好的稳定性。这也证明了该算法在处理不平衡数据分类问题的可行性和有效性。

为了扩大本文算法对不平衡数据分类的适用性，后续研究将考虑多分类问题及相应的评价指标，并进一步改善本文算法对不平衡数据的分类性能。

参考文献：

[1] Laurikkala J. Improving identification of difficult small classes by balancing class distribution [C]// Proc of the 8th Conference on AI in Medicine. 2001: 63-66.

[2] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique [J]. Artificial Intelligence Research, 2002, 16 (3): 321-357.

[3] 郑文昌，陈淑燕，王宣强．面向不平衡数据集的 SMOTE-SVM 交通事件检测算法 [J]．武汉理工大学学报, 2012, 34 (11): 58-62+123.

[4] 衣柏衡，朱建军，李杰．基于改进 SMOTE 的小额贷款公司客户信用风险非均衡 SVM 分类 [J]．中国管理科学, 2016, 24 (03): 24-30.

[5] 杨毅，卢诚波，徐根海．面向不平衡数据集的一种精化 Borderline-SMOTE 方法 [J]．复旦学报：自然科学版, 2017, 56 (05): 537-544.

[6] 蒋盛益，谢照青，余雯．基于代价敏感的朴素贝叶斯不平衡数据分类研究 [J]．计算机研究与发展, 2011, (S1): 387-390.

[7] 李勇，刘战东，张海军．不平衡数据的集成分类算法综述 [J]．计算机应用研究, 2014, 31 (5): 1287-1291.

[8] 邹鹏，莫佳卉，江亦华，等．基于代价敏感决策树的客户价值细分 [J].

管理科学, 2011, 24 (2): 20-29.
[9] 师彦文，王宏杰．基于新型不纯度度量的代价敏感随机森林分类器 [J]. 计算机科学, 2017, 44 (S2): 98-101.
[10] Schapire R E. The strength of weak learnability [C]// Proc of the 2nd Annual Workshop on Computational Learning Theory. 1989: 197-227.
[11] Breiman L. Bagging predictors [J]. Machine Learning, 1996, 24 (2): 123-140.
[12] Li K, Fang X, Zhai J, et al. An imbalanced data classification method driven by boundary samples-boundary-boost [C]// Proc of International Conference on Information Science and Control Engineering. 2016: 194-199.
[13] 胡小生，温菊屏，钟勇．动态平衡采样的不平衡数据集成分类方法 [J]. 智能系统学报, 2016, 11 (02): 257-263.
[14] 秦孟梅，邱建林，陆鹏程，等．基于AdaBoost的类不平衡学习算法 [J]. 计算机应用研究, 2017, 34 (11): 3229-3232+3254.
[15] 应维云，蔺楠，谢雅雅，等．用LDA Boosting算法进行客户流失预测 [J]. 数理统计与管理, 2010 (3): 400-408.
[16] 李诒靖，郭海湘，李亚楠，等．一种基于Boosting的集成学习算法在不均衡数据中的分类 [J]. 系统工程理论与实践, 2016 (1): 189-199.
[17] 王璐林．面向不平衡样本的Boosting分类算法研究 [D]. 哈尔滨：哈尔滨工业大学, 2013.
[18] 李想. Boosting分类算法的应用与研究 [D]. 兰州：兰州交通大学, 2012.
[19] 常志朋，程龙生．核Fisher判别分析多参数自动优化算法 [J]. 系统工程与电子技术, 2013, (01): 212-217.
[20] 李建云，邱菀华．核Fisher判别分析方法评估消费者信用风险 [J]. 系统工程理论方法应用, 2004 (6): 548-552+556.